Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning

Hui Han[1], Wen-Yuan Wang[1], and Bing-Huan Mao[2]

[1] Department of Automation, Tsinghua University, Beijing 100084, P. R. China
hanh01@mails.tsinghua.edu.cn
wwy-dau@mail.tsinghua.edu.cn
[2] Department of Statistics, Central University of Finance and Economics, Beijing 100081, P. R. China
maobinghuan@yahoo.com

Abstract. In recent years, mining with imbalanced data sets receives more and more attentions in both theoretical and practical aspects. This paper introduces the importance of imbalanced data sets and their broad application domains in data mining, and then summarizes the evaluation metrics and the existing methods to evaluate and solve the imbalance problem. Synthetic minority over-sampling technique (SMOTE) is one of the over-sampling methods addressing this problem. Based on SMOTE method, this paper presents two new minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, in which only the minority examples near the borderline are over-sampled. For the minority class, experiments show that our approaches achieve better TP rate and F-value than SMOTE and random over-sampling methods.

1 Introduction

There may be two kinds of imbalances in a data set. One is between-class imbalance, in which case some classes have much more examples than others [1]. The other is within-class imbalance, in which case some subsets of one class have much fewer examples than other subsets of the same class [2]. By convention, in imbalanced data sets, we call the classes having more examples the majority classes and the ones having fewer examples the minority classes.

The problem of imbalance has got more and more emphasis in recent years. Imbalanced data sets exists in many real-world domains, such as spotting unreliable telecommunication customers [3], detection of oil spills in satellite radar images [4], learning word pronunciations [5], text classification [6], detection of fraudulent telephone calls [7], information retrieval and filtering tasks [8], and so on. In these domains, what we are really interested in is the minority class other than the majority class. Thus, we need a fairly high prediction for the minority class. However, the traditional data mining algorithms behaves undesirable in the instance of imbalanced data sets, as the distribution of the data sets is not taken into consideration when these algorithms are designed.

The structure of this paper is organized as follows. Section 2 gives a brief introduction to the recent developments in the domains of imbalanced data sets. Section 3

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part I, LNCS 3644, pp. 878 – 887, 2005.

describes our over-sampling methods on resolving the imbalanced problem. Section 4 presents the experiments and compares our methods with other over-sampling methods. Section 5 draws the conclusion.

2. The Recent Developments in Imbalanced Data Sets Learning

2.1 Evaluation Metrics in Imbalanced Domains

Most of the studies in imbalanced domains mainly concentrate on two-class problem as multi-class problem can be simplified to two-class problem. By convention, the class label of the minority class is positive, and the class label of the majority class is negative. Table 1 illustrates a confusion matrix of a two-class problem. The first column of the table is the actual class label of the examples, and the first row presents their predicted class label. TP and TN denote the number of positive and negative examples that are classified correctly, while FN and FP denote the number of misclassified positive and negative examples respectively.

Table 1. Confusion matrix for a two-class problem

	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{FP}+\text{TN}) \tag{1}$$

$$\text{FP rate} = \text{FP}/(\text{TN}+\text{FP}) \tag{2}$$

$$\text{TP rate} = \text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{3}$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{4}$$

$$F-value = ((1+\beta^2)\cdot \text{Re}call \cdot \text{Pr}ecision)/(\beta^2 \cdot \text{Re}call + \text{Pr}ecision) \tag{5}$$

When used to evaluate the performance of a learner for imbalanced data sets, accuracy is generally apt to predict the majority class better and behaves poorly to the minority class. We can come to this conclusion from its definition (formula (1)): if the dataset is extremely imbalanced, even when the classifier classifies all the majority examples correctly and misclassifies all the minority examples, the accuracy of the learner is still high because there are much more majority examples than minority examples. Under the circumstance, accuracy can not reflect reliable prediction for the minority class. Thus, more reasonable evaluation metrics are needed.

ROC curve [9] is one of the popular metrics to evaluate the learners for imbalanced data sets. It is a two-dimensional graph in which TP rate is plotted on the y-axis and

FP rate is plotted on the x-axis. FP rate (formula (2)) denotes the percentage of the misclassified negative examples, and TP rate (formula (3)) is the percentage of the correctly classified positive examples. The point (0, 1) is the ideal point of the learners. ROC curve depicts relative trade-offs between benefits (TP rate) and costs (FP rate). AUC (Area under ROC) can also be applied to evaluate the imbalanced data sets [9]. Furthermore, F-value (formula (5)) is also a popular evaluation metric for imbalance problem [10]. It is a kind of combination of recall (formula (3)) and precision (formula (4)), which are effective metrics for information retrieval community where the imbalance problem exist. F-value is high when both recall and precision are high, and can be adjusted through changing the value of β, where β corresponds to relative importance of precision vs. recall and it is usually set to 1.

The above evaluation metrics can reasonably evaluate the learner for imbalanced data sets because their formulae are relative to the minority class.

2.2 Methods for Dealing with Imbalanced Data Sets Learning

The solutions to imbalanced data sets can be divided into data and algorithmic levels. categories. The methods at data level change the distribution of the imbalanced data sets, and then the balanced data sets are provided to the learner to improve the detection rate of minority class. The methods at the algorithm level modify the existing data mining algorithms or put forward new algorithms to resolve the imbalance problem.

2.2.1 The Methods at Data Level

At the data level, different forms of re-sampling methods were proposed [1]. The simplest re-sampling methods are random over-sampling and random under-sampling. The former augments the minority class by exactly duplicating the examples of the minority class, while the latter randomly takes away some examples of the majority class. However, random over-sampling may make the decision regions of the learner smaller and more specific, thus cause the learner to over-fit. Random under-sampling can reduce some useful information of the data sets. Many improved re-sampling methods are thus presented, such as heuristic re-sampling methods, combination of over-sampling and under-sampling methods, embedding re-sampling methods into data mining algorithms, and so on. Some of the improved re-sampling methods are as follows.

Kubat et al. presented a heuristic under-sampling method which balanced the data set through eliminating the noise and redundant examples of the majority class [11]. Nitesh et al. over-sampled the minority class through SMOTE (Synthetic Minority Over-sampling Technique) method, which generated new synthetic examples along the line between the minority examples and their selected nearest neighbors [12]. The advantage of SMOTE is that it makes the decision regions larger and less specific. Nitesh et al. integrated SMOTE into a standard boosting procedure, thus improved the prediction of the minority class while not sacrificing the accuracy of the whole testing set [13]. Gustavo et al. combined over-sampling and under-sampling methods to resolve the imbalanced problem [14]. Andrew Estabrooks et al. proposed a multiple re-sampling method which selected the most appropriate re-sampling rate adaptively [15]. Taeho Jo et al. put forward a cluster-based over-sampling method which dealt

with between-class imbalance and within-class imbalance simultaneously [16]. Hongyu Guo et al. found out hard examples of the majority and minority classes during the process of boosting, then generated new synthetic examples from hard examples and add them to the data sets [17].

2.2.2 The Methods at Algorithm Level

The methods at algorithm level operate on the algorithms other than the data sets. The standard boosting algorithm, e.g. Adaboost [18], increases the weights of misclassified examples and decreases those correctly classified using the same proportion, without considering the imbalance of the data sets. Thus, traditional boosting algorithms do not perform well on the minority class. Aiming at the disadvantage above, Mahesh V. Joshi et al. proposed an improved boosting algorithm which updated weights of positive prediction (TP and FP) differently from weights of negative prediction (TN and FN). The new algorithm can achieve better prediction for the minority class [19]. When dealing with imbalanced data sets, the class boundary learned by Support Vector Machines (SVMs) is apt to skew toward the minority class, thus increase the misclassified rate of the minority class. Gang Wu et al. proposed class-boundary alignment algorithm which modify the class boundary through changing the kernel function of SVMs [20]. Kaizhu Huang et al. presented Biased Minimax Probability Machine (BMPM) to resolve the imbalance problem. Given the reliable mean and covariance matrices of the majority and minority classes, BMPM can derive the decision hyperplane by adjusting the lower bound of the real accuracy of the testing set [21]. Furthermore, there are other effective methods such as cost-based learning, adjusting the probability of the learners and one-class learning, and so on [22] [23].

3 A New Over-Sampling Method: Borderline-SMOTE

In order to achieve better prediction, most of the classification algorithms attempt to learn the borderline of each class as exactly as possible in the training process. The examples on the borderline and the ones nearby (we call them borderline examples in this paper) are more apt to be misclassified than the ones far from the borderline, and thus more important for classification.

Based on the analysis above, those examples far from the borderline may contribute little to classification. We thus present two new minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, in which only the borderline examples of the minority class are over-sampled. Our methods are different from the existing over-sampling methods in which all the minority examples or a random subset of the minority class are over-sampled [1] [2] [12].

Our methods are based on SMOTE (Synthetic Minority Over-sampling Technique) [12]. SMOTE generates synthetic minority examples to over-sample the minority class. For every minority example, its k (which is set to 5 in SMOTE) nearest neighbors of the same class are calculated, then some examples are randomly selected from them according to the over-sampling rate. After that, new synthetic examples are generated along the line between the minority example and its selected nearest neighbors. Not like the existing over-sampling methods, our methods only over-sample or strengthen the borderline minority examples. First, we find out the border-

line minority examples; then, synthetic examples are generated from them and added to the original training set. Suppose that the whole training set is *T*, the minority class is *P* and the majority class is *N*, and

$$P = \{p_1, p_2, ..., p_{pnum}\}, N = \{n_1, n_2, ..., n_{nnum}\}$$

where *pnum* and *nnum* are the number of minority and majority examples. The detailed procedure of borderline-SMOTE1 is as follows.

Step 1. For every $p_i (i = 1,2,..., pnum)$ *in the minority class P, we calculate its m nearest neighbors from the whole training set T. The number of majority examples among the m nearest neighbors is denoted by* $m' (0 \le m' \le m)$.

Step 2. If $m' = m$, *i.e. all the m nearest neighbors of* p_i *are majority examples,* p_i *is considered to be noise and is not operated in the following steps. If* $m/2 \le m' < m$, *namely the number of* p_i*'s majority nearest neighbors is larger than the number of its minority ones,* p_i *is considered to be easily misclassified and put into a set DANGER. If* $0 \le m' < m/2$, p_i *is safe and needs not to participate in the follows steps.*

Step 3. The examples in DANGER are the borderline data of the minority class P, and we can see that $DANGER \subseteq P$. *We set*

$$DANGER = \{p'_1, p'_2, ..., p'_{dnum}\}, \quad 0 \le dnum \le pnum$$

For each example in *DANGER*, we calculate its *k* nearest neighbors from *P*.

Step 4. In this step, we generate $s \times dnum$ *synthetic positive examples from the data in DANGER, where s is an integer between 1 and k. For each* p'_i, *we randomly select s nearest neighbors from its k nearest neighbors in P. Firstly, we calculate the differences,* dif_j $(j = 1,2,..., s)$ *between* p'_i *and its s nearest neighbors from P, then multiply* dif_j *by a random number* r_j $(j = 1,2,..., s)$ *between 0 and 1, finally, s new synthetic minority examples are generated between* p'_i *and its nearest neighbors:*

$$synthetic_j = p'_i + r_j \times dif_j, \quad j = 1,2,..., s$$

We repeat the above procedure for each p'_i in *DANGER* and can attain $s \times dnum$ synthetic examples. This step is similar with SMOTE, for more detail see [12].

In the procedure above, p_i, n_i, p'_i, dif_j and $synthetic_j$ are vectors. We can see that new synthetic data are generated along the line between the minority borderline examples and their nearest neighbors of the same class, thus strengthened the borderline examples.

Borderline-SMOTE2 not only generates synthetic examples from each example in *DANGER* and its positive nearest neighbors in *P*, but also does that from its nearest negative neighbor in *N*. The difference between it and its nearest negative neighbor is

multiplied a random number between 0 and 0.5, thus the new generated examples are closer to the minority class.

Our methods can be easily understood with the following simulated data set, Circle, which has two classes. Fig. 1 (a) shows the original distribution of the data set, the circle points represent majority examples and the plus signs are minority examples. Firstly, we apply borderline-SMOTE to find out the borderline examples of the minority class, which are denoted by solid squares in Fig. 1 (b). Then, new synthetic examples are generated through those borderline examples of the minority class. The synthetic examples are shown in Fig. 1 (c) with hollow squares. It is easy to find out from the figures that, different from SMOTE, our methods only over-sample or strengthen the borderline and its nearby points of the minority class.

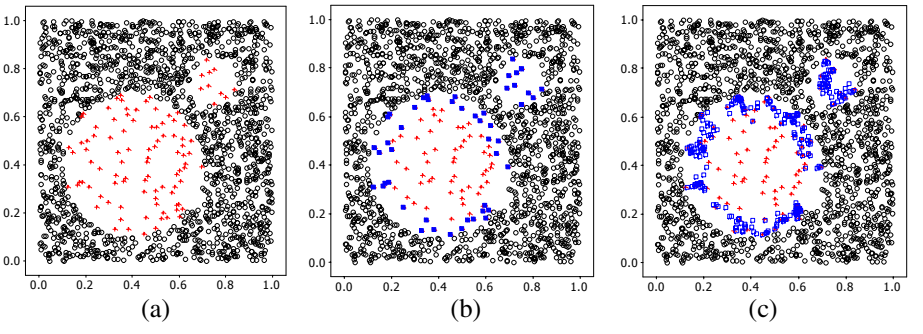

Fig. 1. (a) The original distribution of Circle data set. (b) The borderline minority examples (*solid squares*). (c) The borderline synthetic minority examples (*hollow squares*).

4 Experiments

We use TP rate and F-value for the minority class to evaluate the results of our experiments. TP rate denotes the accuracy of the minority class. And the value of β in F-value is set to 1 in this paper.

The four data sets used in our experiments are shown in Table 2. Among the four data sets, Circle is our simulated data set depicted in Fig. 1, and the others are from UCI [24]. All the attributes in the data sets are quantitative. For Satimage, we choose class label "4" as the minority class and regard the remainders as the majority class, as we only study two-class problem in this paper.

Table 2. The description of the data sets

The name of Data set	number of Examples	number of Attributes	Class label (minority : majority)	Percentage of minority class
Circle(Simulation)	1600	2	1:0	6.25%
Pima(UCI)	768	8	1:0	34.77%
Satimage(UCI)	6435	36	4:remainder	9.73%
Haberman(UCI)	306	3	Die : Survive	26.47%

In our experiments, four over-sampling methods are applied to the data sets: SMOTE, random over-sampling and our methods, borderline-SMOTE1 and borderline-SMOTE2, among which random over-sampling method augments the minority class by exactly duplicating the positive examples partly or completely [1]. Through increasing the number of examples in the minority class, over-sampling methods can balance the distribution of the data sets and improve the detection rate of the minority class.

In order to compare the results conveniently, the value of m in our methods is set in a way that, the number of the minority examples in *DANGER* is about half of the minority class. The value of k is set to 5 like SMOTE. For each method, the TP rates and F-values are attained through 10-fold cross-validation. In order to decrease the randomness in SMOTE and our methods, the TP rates and F-values for these methods are the average results of three independent 10-fold cross-validation experiments. After the original training sets are over-sampled with the methods above, C4.5 is applied as the validation classifier [25].

Since the nature of imbalance problem is to improve the prediction performance of the minority class, we only present the results of the minority class. We compare the results of the data sets through TP rate and F-value of the minority class. TP rate reflects the performance of the learner on the minority class of the testing set, while F-value shows the performance of the learner on the whole testing set.

Fig. 2 shows our experimental results. In the figure, (a), (b), (c) and (d) depict the F-value and TP rate for the minority class when the four over-sampling methods are applied on Circle, Pima, Satimage and Haberman respectively. The x-axis in each figure is the number of the new synthetic examples. The F-value and TP rate of the original data sets with C4.5 are also shown in the figures.

The results illustrated in Fig. 2 reveal the following results. First of all, all the four over-sampling methods improve TP rate of the minority class. For Circle, Pima and Haberman, the TP rates of our methods are better than SMOTE and random over-sampling. Comparing with the original data sets, the best improvements of TP rate for borderline-SMOTE1 and borderline-SMOTE2 on Circle are 20 and 22 per cent, 21.3 and 20.5 per cent on Pima, 10.1 and 10.0 per cent on Satimage, and both 45.2 per cent on Haberman. For Satimage, the TP rates of our methods are lager than that of random over-sampling, and are comparable with SMOTE. Secondly, the F-value of borderline-SMOTE1 is generally better than SMOTE and random over-sampling, and the F-value of borderline-SMOTE2 is also comparable with others. Comparing with the original data sets, the best improvements of F-value for borderline-SMOTE1 and borderline-SMOTE2 on Circle are 12.1 and 10.3 per cent, 2.3 and 1.3 per cent on Pima, 2.3 and 1.4 per cent on Satimage, and 24.7 and 23.0 per cent on Haberman.

As a whole, border-SMOTE1 behaves excellent on both TP rate and F-value, and borderline-SMOTE2 behaves super on TP rate because it generates synthetic examples from both the minority borderline examples and their nearest neighbors of the majority class, however, the procedure causes overlap between the two classes, thus decreases its F-value to some extent.

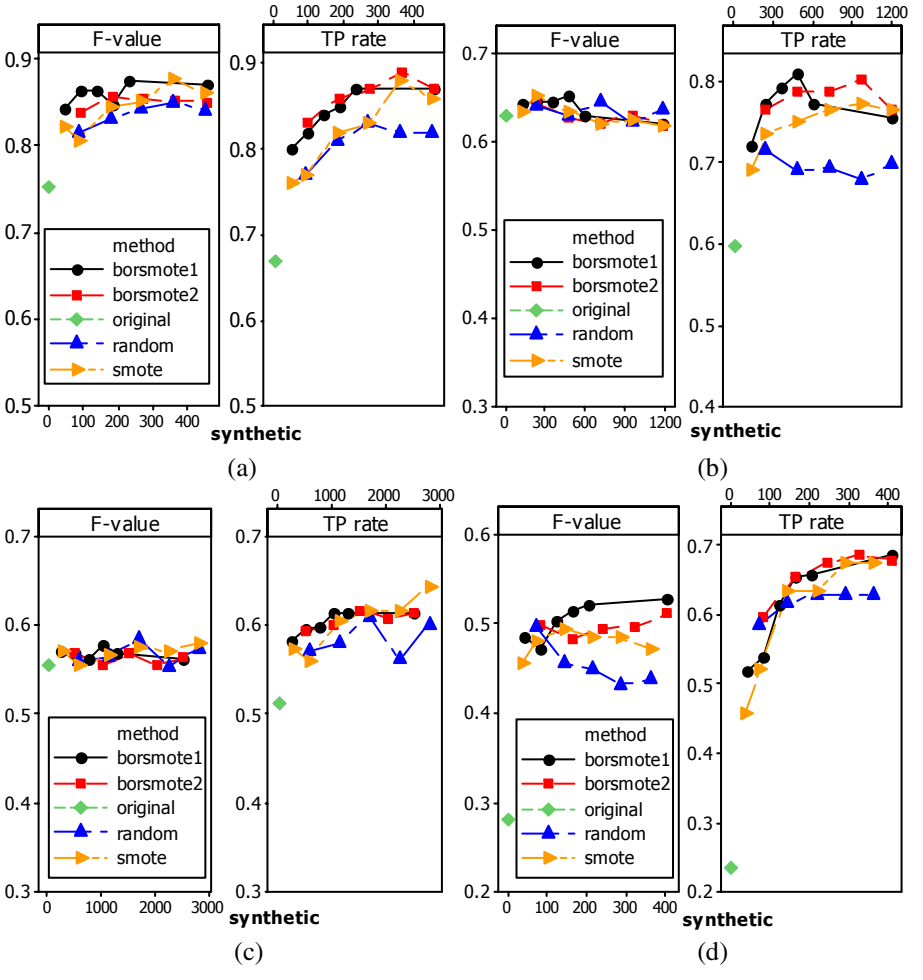

Fig. 2. (a), (b), (c) and (d) illustrate the F-value and TP rate for minority class when proposed over-sampling methods are applied on Circle, Pima, Satimge and Haberman respectively with C4.5. "borsmote1" and "borsmote2" denote borderline-SMOTE1 and borderline-SMOTE2, "random" denotes random over-sampling, and "original" denotes the values of the original data sets. The x-axis is the number of synthetic examples

5 Conclusion

In recent years, learning with imbalanced data sets receives more and more attentions in both theoretical and practical aspects. However, traditional data mining methods are not satisfactory. Aiming to solve the problem, two new synthetic minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2 are presented in this paper. We compared the TP rate and F-value of our methods with SMOTE, random over-sampling and the original C4.5 for four data sets.

The borderline examples of the minority class are more easily misclassified than those ones far from the borderline. Thus our methods only over-sample the borderline examples of the minority class, while SMOTE and random over-sampling augment the minority class through all the examples from the minority class or a random subset of the minority class. Experiments indicate that our methods behave better, which validates the efficiency of our methods.

There are several topics left to be considered further in this line of research. Different strategies to define the *DANGER* examples, and automated adaptive determination of the number of examples in *DANGER* would be valuable. The combination of our methods with under-sampling methods and the integration of our methods to some data mining algorithms, are also worth trying.

References

1. Nitesh V.Chawla, Nathalie Japkowicz and Aleksander Kolcz.: Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explorations 6 (1) (2004) 1-6
2. G. Weiss: Mining with rarity: A unifying framework. SIGKDD Explorations 6 (1) (2004) 7-19
3. Ezawa, K.J., Singh, M. and Norton, S.W.: Learning Goal Oriented Bayesian Networks for Telecommunications Management. In Proceedings of the International Conference on Machine Learning, ICML'96(pp. 139-147), Bari, Italy, Morgan Kaufmann (1996)
4. Kubat, m., Holte, R., and Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning 30 195-215
5. A. van den Bosch, T. Weijters, H. J. van den Herik, and W. Daelemans: When small disjuncts abound, try lazy learning: A case study. In Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning (1997) 109-118
6. Zhaohui Zheng, Xiaoyun Wu, Rohini Srihari: Feature Selection for Text Categorization on Imbalanced Data. SIGKDD Explorations 6 (1) (2004) 80-89
7. Fawcett, T.and Provost, F.: Combining Data Mining and Machine Learning for Effective User Profile. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland OR, AAAI Press (1996) 8-13
8. Lewis, D. and Catlett, Heterogeneous, J.: Uncertainty Sampling for Supervized Learning. Proceedings of the 11th International Conference on Machine Learning, ICML'94 (1994) 148-156
9. Bradley A.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30 (7) (1997) 1145-1159
10. Rijsbergen, C. J. van: Information Retrieval, Butterworths, London (1979)
11. Kubat, M., and Matwin, S. Addressing the Course of Imbalanced Training Sets: One-sided Selection. In ICML'97 (1997) 179-186
12. Chawla, N.V., Bowyer,K.W., Hall, L.O., Kegelmeyer W.P.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16 (2002) 321-357
13. Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, K.: SMOTEBoost: Improving prediction of the Minority Class in Boosting. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat Dubrovnik, Croatia (2003) 107-119
14. Gustavo, E.A., Batista, P.A., Ronaldo, C., Prati, Maria Carolina Monard: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explorations 6 (1) (2004) 20-29

15. Andrew Estabrooks, Taeho Jo and Nathalie Japkowicz: A Multiple Resampling Method for Learning from Imbalanced Data Sets. Comprtational Intelligence 20 (1) (2004) 18-36
16. Taeho Jo, Nathalie Japkowicz: Class Imbalances versus Small Disjuncts. Sigkdd Explorations 6 (1) (2004) 40-49
17. Hongyu Guo, Herna L Viktor: Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. Sigkdd Explorations 6 (1) (2004) 30-39
18. Yoav Freund, Robert Schapire: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1) (1997) 119-139
19. Joshi, M., Kumar,V., Agarwal, R.: Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. First IEEE International Conference on Data Mining, San Jose, CA (2001)
20. Gang Wu, Edward Y.Chang. Class-Boundary Alignment for Imbalanced Dataset Learning. Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC (2003)
21. Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu. Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004)
22. Dietterich, T., Margineantu, D., Provost, F. and P. Turney, edited, Proceedings of the ICML'2000 Workshop on Cost-sensitive Learning (2000)
23. Manevitz, L.M. and Yousef, M.: One-class SVMs for document classification. Journal of Machine Learning Research 2 (2001) 139-154
24. Blake, C., & Merz, C. (1998). UCI Repository of Machine Learning Databases http://www.ics.uci.edu/ ~ mlearn/ ~ MLRepository.html. Department of Information and Computer Sciences, University of California, Irvine
25. Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1992)